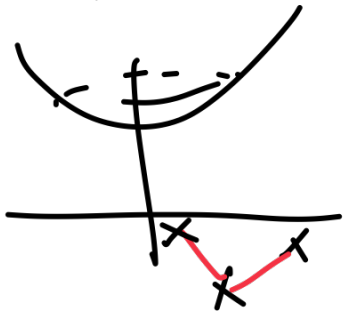
Linear regression $\quad x \in \mathbb{R}^D$

$$h_\beta(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_D x_D$$

$\{x^{(i)}, t^{(i)}\}_{i=1}^{N}$

$$\rightarrow \ell(\beta) = \min_\beta \sum_{i=1}^{N} \left(t^{(i)} - h_\beta(x^{(i)})\right)^2$$

$\rightarrow$ approach : gradient descent



$$\beta^{(0)} \qquad \beta^{(k+1)} \leftarrow \beta^{(k)} - \eta \operatorname*{grad}_\beta \ell(\beta)$$

possible extension to datasets not linearly distributed

$$\text{gradient} = \left(\frac{\partial \ell}{\partial \beta_0}, \frac{\partial \ell}{\partial \beta_1}, \dots \frac{\partial \ell}{\partial \beta_D} \right) \qquad X = (x_1, x_2, \dots, x_D) \in \mathbb{R}^D$$

$$h_\beta(x^{(i)}) = \beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)}$$

$$= \langle \vec{\beta}, (1, \vec{x}) \rangle \qquad \tilde{x} = [1, \vec{x}]$$

$$\tilde{x} = [1, x_1, x_2, \dots, x_D]$$

$$h_\beta(x^{(i)}) = \langle \vec{\beta}, \tilde{x} \rangle$$

$$= \vec{\beta}^T \tilde{x} = [\beta_0, \beta_1, \dots \beta_D] \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_D \end{bmatrix}$$

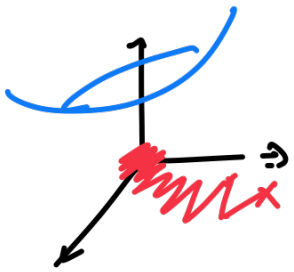
$$\min_\beta \frac{1}{N} \sum_{i \in \mathcal{D}} (t^{(i)} - h_\beta(x^{(i)}))^2$$

Batch gradient descent

(I) $$\beta^{(k+1)} \leftarrow \beta^{(k)} - \eta \operatorname{grad}_\beta \ell_{BGD}(\beta)$$

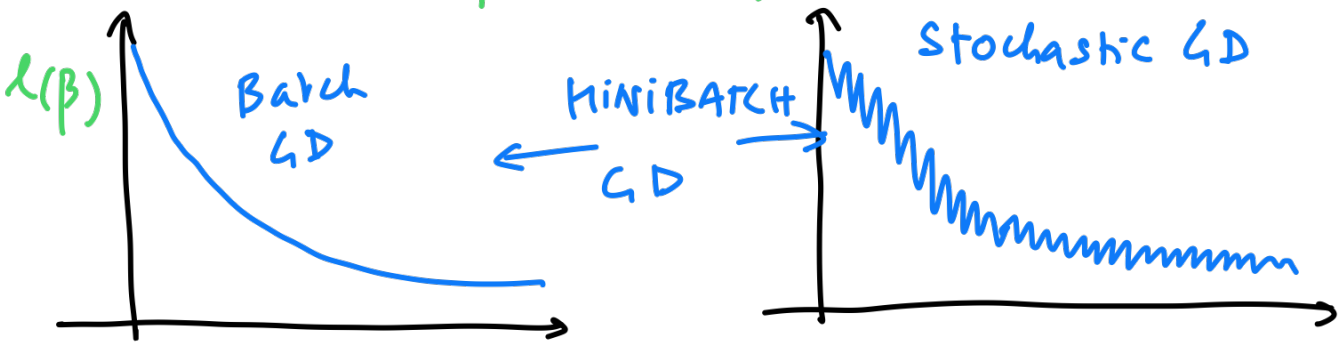
(II) for $i \in \mathcal{D}$

$$\min_\beta (t^{(i)} - h_\beta(x^{(i)}))^2$$



$$\beta^{(k+1)} \leftarrow \beta^{(k)} - \eta \operatorname{grad} \ell_{SGD}(\beta)$$

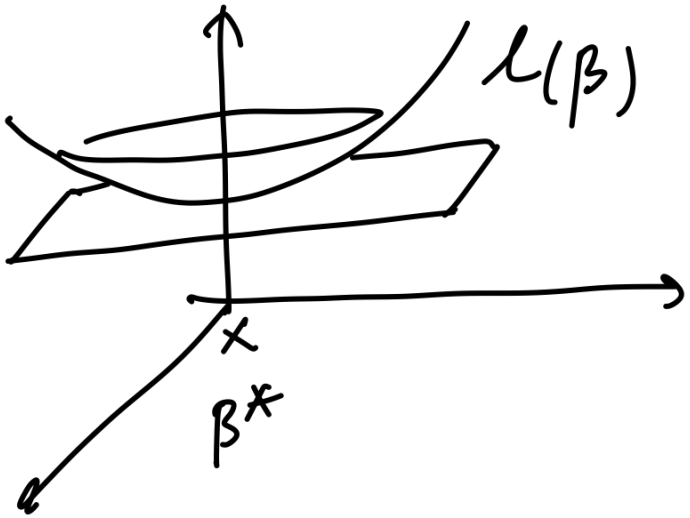
Stochastic GD



by incorporating polynomial features.

Today: approach #2: Normal equations

statistical intuition



$$\underset{\sim}{\underset{=}{X}} = \begin{bmatrix} — & \tilde{x}^{(1)} & — \\ — & \tilde{x}^{(2)} & — \\ & \vdots & \\ — & \tilde{x}^{(N)} & — \end{bmatrix}$$

$$\vec{t} = \begin{bmatrix} t^{(1)} \\ \vdots \\ t^{(N)} \end{bmatrix} \qquad \vec{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_D \end{bmatrix}$$

$$\vec{e} = \vec{t} - \underset{\sim}{\underset{=}{X}}\vec{\beta} \qquad \|\vec{e}\|^2 = e^T e = \langle e, e \rangle$$

$$e = (e^{(1)}, e^{(2)}, \cdots, e^{(N)}) \qquad = \sum_{i=1}^{N} (e^{(i)})^2$$

$$e^T e = \sum_{i=2}^{N} (e^{(i)})^2 = \sum_{i=2}^{N} (t^{(i)} - (\tilde{x}^{(i)})^T \beta)^2 = l(\beta)$$

$$l(\beta) = e^T e = (\vec{t} - \underline{\underline{\tilde{X}}}\,\vec{\beta})^T (\vec{t} - \underline{\underline{\tilde{X}}}\,\vec{\beta}) \frac{1}{N}$$

$$l(\beta) = (\vec{t}^{\,T}\vec{t} - \vec{\beta}^T \underline{\underline{\tilde{X}}}^T \vec{t} - \vec{t}^{\,T} \underline{\underline{\tilde{X}}}\,\vec{\beta} + \beta^T \underline{\underline{\tilde{X}}}^T \underline{\underline{\tilde{X}}}\,\beta) \frac{1}{N}$$

$$-2\,\underline{\underline{\tilde{X}}}^T \vec{t} \qquad\qquad 2\,\underline{\underline{\tilde{X}}}^T \underline{\underline{\tilde{X}}}\,\beta$$

$$\operatorname*{grad}_{\beta} \beta^T A \beta = \operatorname*{grad}_{\beta} [\beta_0, \beta_1] \begin{bmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = 2A\beta$$

$$\underset{\beta}{\text{grad}} \quad \beta_0^2 a_{11} + \beta_1^2 a_{22} + \beta_0 \beta_1 a_{12} + \beta_1 \beta_0 a_{12}$$

$$= \beta_0^2 a_{11} + \beta_1^2 a_{22} + 2\beta_0 \beta_1 a_{12}$$

$$\underset{\beta}{\text{grad}} = \left[\frac{\partial \ell}{\partial \beta_0}, \frac{\partial \ell}{\partial \beta_1} \right] = \left[2\beta_0 a_{11} + 2\beta_1 a_{12}, \; 2\beta_1 a_{22} + 2\beta_0 a_{12} \right]$$

$$2 \begin{bmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

$$\underset{\beta}{\text{grad}} \; \ell(\beta) = -2 \underline{\underline{\tilde{X}}}^T t + 2 \underline{\underline{\tilde{X}}}^T \underline{\underline{\tilde{X}}} \beta = 0$$

$$2 \underline{\underline{\tilde{X}}}^T \underline{\underline{\tilde{X}}} \beta = 2 \underline{\underline{\tilde{X}}}^T t \quad \rightarrow \text{Normal equations}$$

$$\begin{bmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{bmatrix}$$

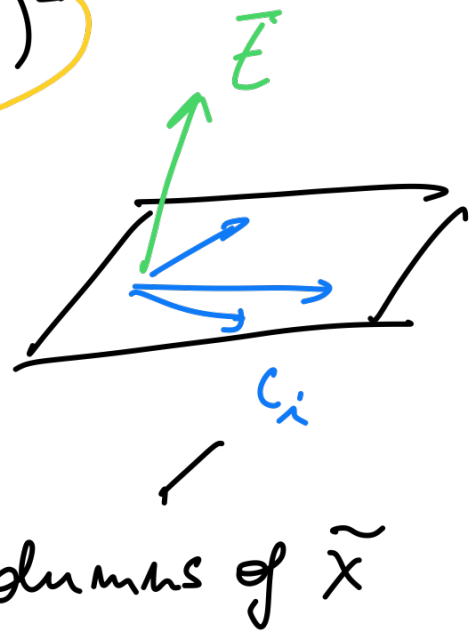
$$\underbrace{(\tilde{X}^T\tilde{X})^{-1}(\tilde{X}^T\tilde{X})}_{I}\beta = (\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T\vec{t}$$

$$\vec{\beta} = (\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T\vec{t}$$

$$\min_{\beta} \|t - \tilde{X}\beta\|_2^2 = \sum_{i=1}^{N} (t^{(i)} - \tilde{x}^{(i)T}\beta)^2$$

$$\tilde{X} = \begin{bmatrix} — & x^{(1)} & — \\ — & x^{(2)} & — \\ & \vdots & \\ — & x^{(N)} & — \end{bmatrix} \qquad \tilde{X}\beta = \vec{t}$$

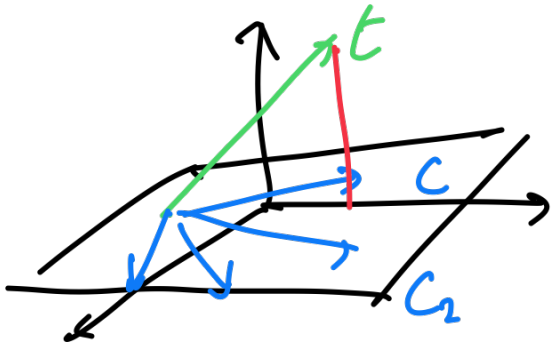
$$\|e\|_2^2 = \sum_{i=1}^{N} (e^{(i)})^2$$



columns of $\tilde{X}$

Condition to use the normal equations $\tilde{X}^T \tilde{X}$ has to be invertible

$$\tilde{X} = \begin{bmatrix} | & | & | & & | \\ c_0 & c_1 & c_2 & \cdots & c_p \\ | & | & | & & | \end{bmatrix}$$



What happens if c_k can be (almost) derived from the other columns ce

When solving $\min_{\beta} \sum_{i=2}^{N} (t^{(i)} - (\tilde{x}^{(i)})^T \beta)^2 = \| \vec{t} - \tilde{X} \vec{\beta} \|_2^2$

there still exists an orthogonal projection but this projection is not uniquely defined from the c_i's

Assume columns G_i's of $\tilde{X}$ are almost linearly dependent

Assume we want to write $\tilde{Z}$ from the first p columns:

$$\langle z_\ell, z_\ell \rangle = z_\ell^T z_\ell$$

Start with $z_0 = G_0$

for $j = 1, \dots, p$

$$\hat{\gamma}_{\ell j} = \langle G_j, z_\ell \rangle / \langle z_\ell, z_\ell \rangle \quad \text{for } \ell = 0, \dots, j-1$$

$$z_j = G_j - \sum_{k=0}^{j-1} \hat{\gamma}_{kj} z_k$$

z_p is the only column that contain column G_p and with a coefficient 1

$\Rightarrow$ as a result $\hat{\beta}_p$ can be obtained by projecting $\vec{E}$ onto z_p

$$\hat{\beta}_p = \frac{\langle z_p, \vec{E} \rangle}{\langle z_p, z_p \rangle} \longrightarrow \ll \varepsilon$$

$$\hat{\beta}_p \gg \varepsilon$$

For large regression coefficients

the model $h_\beta(x) = \beta_0 + \beta_1 x_1 + \cdots + \hat{\beta}_p x_p + \cdots$

applied to a new x will result in inaccuracies (small changes in x_p will lead

to very different predictions)

→ Next Wednesday

→ alternative approaches

→ Best subset selection

→ Ridge Regression

→ LASSO